OF URFS AND ORFS

A Primer on How to Analyze Derived Amino Acid Sequences

RUSSELL F. DOOLITTLE

University Science Books

First published April, 1987

Library of Congress Catalog Card Number: 87-050024

ISBN 0-935702-54-7

Printed in the United States of America

10 9 8 7 6 5 4 3 2

University Science Books
20 Edgehill Road
Mill Valley, CA 94941

ties in a row somewhere. Accordingly, only segments meeting that condition are considered, thereby cutting down not only the number of segments that warrant more serious consideration, but also eliminating much of the time consumed in the more exhaustive overlapping-segment prescreens. Speed is seldom the highest priority for the occasional searcher, however, and it must be kept in mind that prescreening usually results in a small but measurable loss of sensitivity.

Deciding About Significance

It is never possible to prove that two sequences are not evolutionarily related; they may just have changed so much since their divergence point that they are no longer recognizably similar. The sequence divergence itself may have been the result of speciation (species divergence), on the one hand, or substitutions after a gene duplication, on the other.

It is important to understand the limits of what a provable match will be. It often surprises the beginner that two random — or totally unrelated — amino acid sequences may turn out to be 10-20% identical when aligned by a typical computer alignment program. The reason is, of course, that alignment programs allow gaps to be incorporated in either sequence if, by so doing, the alignment is significantly improved. In this regard, an appropriate gap penalty is imposed in order that gapping not get out of hand. Since the computer is allowed to put gaps in sequences whether or not they are evolutionarily related, the percent identity of unrelated proteins will be higher than the 5 or 6% that would be expected if a rigid, gapless scoring system was imposed. As in all probabilistic situations, there will be a dispersion about the mean, and depending on the sequence lengths, two random sequences could be 10-20% identical. What counts, of course, is the overall alignment score; but percent identities are so universally familiar they are useful in setting guidelines.

The Nature of Sequence Divergence

Two diverging sequences take the course of a negative exponential. This follows strictly from classical statistics and the fact that any position can be subject to reverse changes ("back-mutations") and

multiple hits (Zuckerkandl & Pauling, 1965). As such the percent identity (or percent difference) for two sequences is not a proportionate indicator of how much change has actually occurred (Figure 2). Two sequences that are 50% different have actually sustained "hits" amounting to 80 changes per 100 residues. The "hits", or surviving point mutations, are referred to in different terms by different authors. Dayhoff and Eck (1968) introduced the curious but pronounceable pseudoacronym "PAM" for "accepted point mutation," and Holmquist *et al* (1972) have used the "REH" unit for "random evolutionary hit." The perhaps overly vague term "actual replacement" ("AR") is used in this booklet. In any case, a protein may endure up to 360 such changes per 100 residues before a point is reached where it is no longer recognizable (Figure 2).

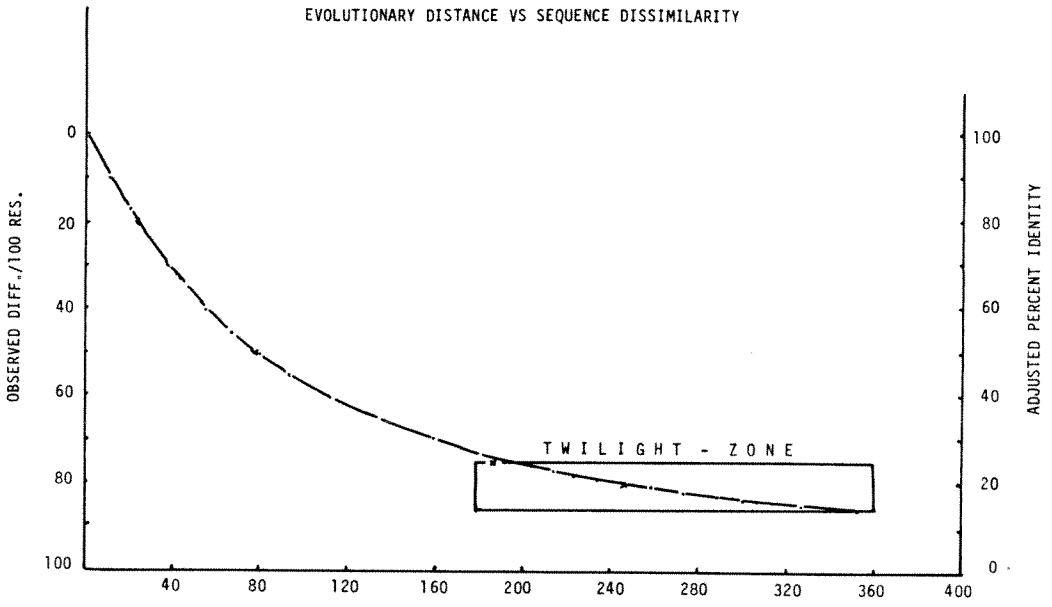

Figure 2. Two randomly diverging sequences change in a negatively exponential fashion. After the insertion of gaps to align two random sequences, it can be expected that they will be 10-20% identical.

Significance: Some Rules of Thumb

At this point, let me offer some "rules of thumb" about degrees of confidence. If two sequences are longer than 100 residues, and are more than 25% identical after suitable gapping, it is very likely they are genuinely related (really, one is saying here that the chances are slim that the similarities are due to chance). If such sequences are 15-25% identical, (the "twilight zone" shown in Figure 2), they may very well be related, and it is worthwhile doing some empirical statistics (described below) to establish a level of confidence. Finally, if the sequences are less than 15% identical, no matter how long they may be, it is going to be difficult to make the point.

The two big variables in comparing amino acid sequences have to do with the lengths of the sequences and their compositions. It is disarmingly simple to find pairs of short sequences that appear similar. Moreover, some proteins are unusually rich in certain amino acids, and this can lead to their appearing more similar than they really are. There is a strategy to correct for both these parameters: it is called the "jumbling" (or "scrambling" or "randomizing") approach. Let us consider two different pairs of sequences to see how it works. One of these pairs, myoglobin and the α chain of hemoglobin, is obviously related (Figure 3, upper); the other, ribonuclease and lysozyme, is sometimes thought to be so (Figure 3, bottom).

First, the two sequences being compared are run through a suitable alignment program and an alignment score obtained. Then both sequences are repeatedly randomized, until a number of jumbled sequences of each are on hand. Then each of the jumbled sequences of one is subjected to the alignment procedure with each of the jumbled versions of the other. For example, if each sequence is jumbled eight times, then altogether 64 alignments of jumbled pairs are made. Their alignment scores are averaged, the mean found and the standard deviation calculated. The score of the authentic alignment is then compared with the mean of the randomized sets, and how much greater (or less) is expressed in standard deviations.

Not everybody has a computer big or fast enough to perform all these jumbled comparisons, of course, and some people may not have access to a computer at all, so some more rules-of-thumb are in order. A simple way of estimating the significance of a match

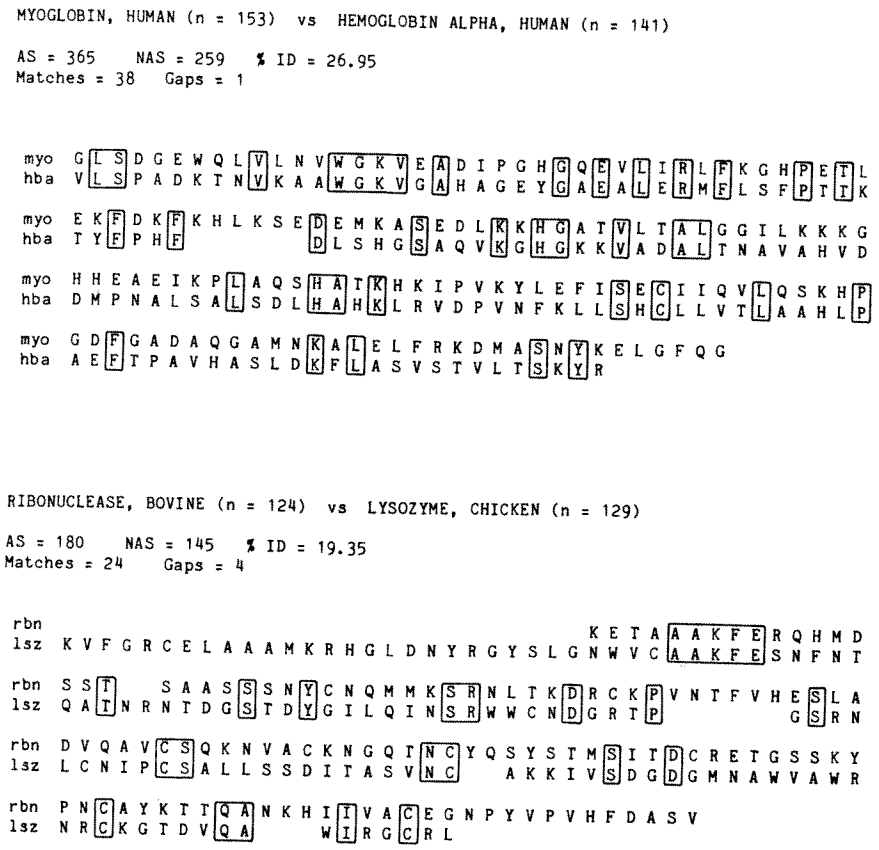

Figure 3. Optimal alignments of human myoglobin and hemoglobin α chain (*upper panel*) and bovine ribonuclease and chicken lysozyme (*lower panel*). Identical residues are boxed.

is as follows. Calculate a simple alignment score by counting every identity as 10 (but 20 for cysteine matches) and subtracting 25 for every gap (Doolittle, 1981). Then normalize the score to 100 residues and consult the graph in Figure 4 that takes account the lengths of the sequences.

Now let us re-examine the two simple alignments in Figure 3. In the first case, human myoglobin is aligned with the α chain of human hemoglobin (upper panel). There are 38 identities, includ-

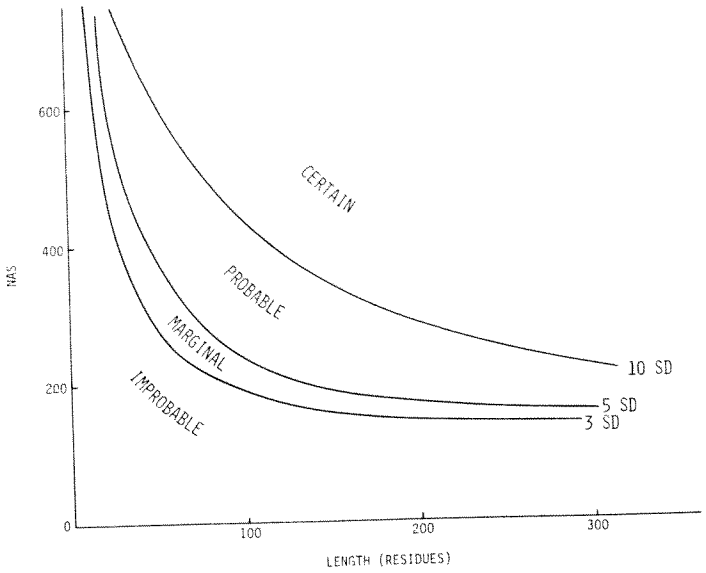

Figure 4. A rough guide to significance in the comparison of amino acid sequences that emphasizes the importance of sequence length. NAS is the "normalized alignment score," in which the number of identities is multiplied by 10 (20 for cysteines) and the number of gaps is multiplied by -25. The score is normalized by dividing by the average length of the two sequences and multiplying by 100 (from Doolittle *et al*, 1986).

ing one cysteine, and only one gap. The alignment score is thus (37 x 10) + (1 x 20) - (1 x 25) = 365. The sequences are 141 and 153 residues in length; the usual convention is to normalize with the shorter length rather than the average, so 365/141 x 100 = 259. When we consult Figure 4, we find that the match is comfortably significant. If we let the computer do the job, we get a much clearer representation, however (Figure 5).

In contrast, when we consider the alignment of bovine pancreatic ribonuclease with chicken eggwhite lysozyme (Figure 3, lower panel) we see that there are only 24 identities, four of which are cysteines, and that there are four gaps:

$$(20 \times 10) + (4 \times 20) - (4 \times 25) = 180$$

$$\frac{180}{124} \times 100 = 145$$

According to Figure 4, this is not a significant match, and when a "jumble" exercise is undertaken, we see that there is no real basis for supposing anything other than a random situation (Figure 5, lower panel).

About Weighted Scales

It is common knowledge that most of the amino-acid replacements in a protein that survive during the course of evolution are structurally conservative (for an extended discussion, see Doolittle, 1979). Can this trend be put to use as a means for sharpening our searches and alignments? To a degree, yes. But probably not as much as believed by some. In our laboratory we undertook a comparison of four different scales to see how effective they were in dealing with distantly related sequences (Feng *et al*, 1985). The four approaches were:

a) identities vs non-identities
b) genetic code base changes (Fitch, 1966)
c) genetic code and structural similarity
d) the Dayhoff Minimum Mutation Matrix (Dayoff, 1978)

In the end, the Minimum Mutation Matrix, which is based on a compilation of observed amino acid replacements between closely related organisms, proved best, but it was closely followed by the method that arbitrarily weighted interchanges on the basis of structural similarities and the ease of genetic interchange (Table III). Neither was overwhelmingly better than the scheme that used only identities, however. The use of the genetic code alone was the least effective. The reason is that of the 75 single-base amino acid interchanges, at least half are between amino acids that are structurally dissimilar.

It must be appreciated that there are exactly 190 interchanges for 20 amino acids:

$$\left(20 \times \frac{19}{2}\right) = 190$$

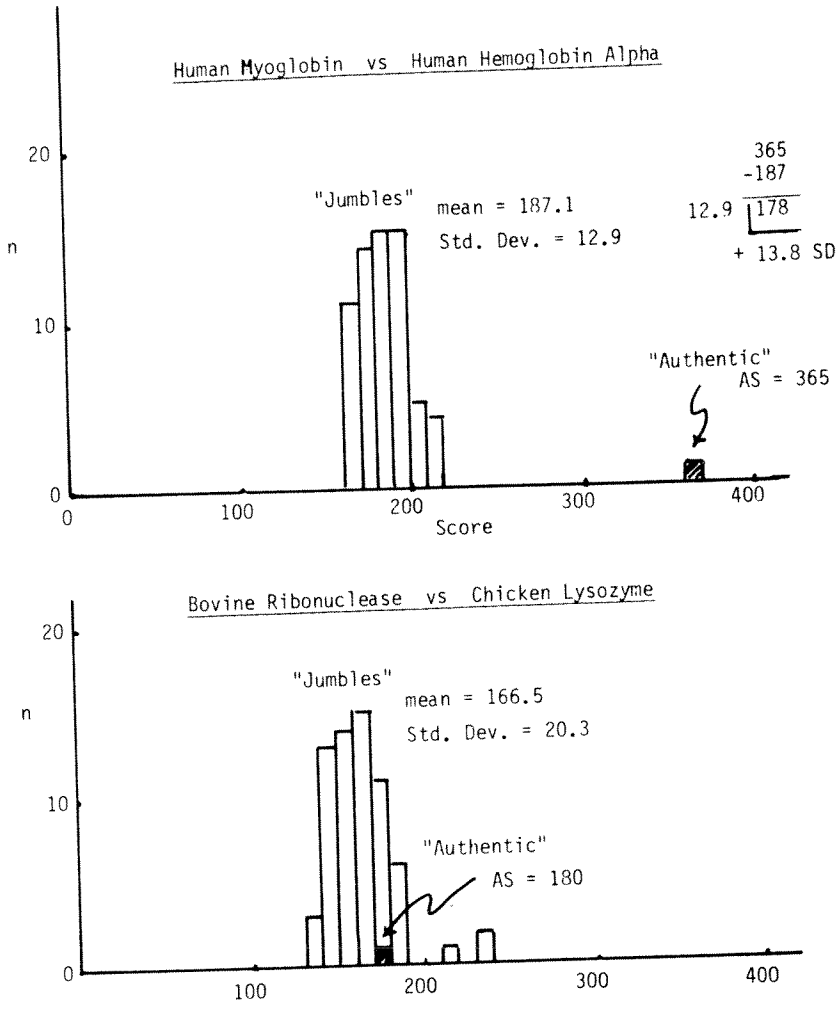

Figure 5. Distribution of alignment scores obtained with jumbled sequences. It is clear that the real comparison of myoglobin and hemoglobin α chain (*upper panel*) had a score that was significantly higher than those obtained with randomized sequences. On the other hand, the real comparison of lysozyme and ribonuclease (*lower panel*) lies amidst those obtained with the randomized sequences of the same lengths and compositions.

Everyone agrees that the changes Glu/Asp, Val/Ile, Ser/Thr, Arg/Lys and Gln/Asn are conservative. The problem arises when one gets past the first dozen or so that are readily categorized. Another consideration to be kept in mind is that identities will always be the most important contributants, even in weighted schemes (highest weight). Further, highly conserved residues may be surrounded by "well-churned" locations that tend to vitiate the value of the weighted scale. Still, on balance, weighted scales are useful devices, and we use them consistently.

Searching Short Sequences

Sometimes a very short sequence is all that's available. Even though it's distressingly easy to match up short peptide sequences, and although it may be difficult to demonstrate significance (Figure 4), it is still worthwhile to search whatever you have. Indeed, sometimes something will turn up that can be directly tested in a laboratory setting. For example, several years ago, Joseph Brown, who was then at the Fred Hutchison Cancer Center in Seattle, purified a melanocyte tumor cell antigen. There was only enough material available for a single microsequencer run, and that endeavor only managed 13 cycles. In fact, only 10 of those first 13 amino acids were identified with any confidence. Still, a search of our data bank turned up only a single candidate: transferrin.

```
Tumor antigen, p97:      G M E V R W C A T S D ? E
                             * * * * *   *     *
Human transferrin:   V P D K K V R W C A V S E H E
```

Brown promptly tested to see if the tumor antigen bound ^{59}Fe. It did, and with the same avidity as transferrin (Brown *et al*, 1982). Today that protein, which appears to be an important factor in rendering the tumor melanocyte immortal, is known to be about 40% identical with the better known transferrins and is called melanotransferrin (Rose *et al*, 1986). The take-home message is: a search of even a very short sequence may put you on the right track and save years of work.

The advent of recombinant DNA technology that allows the production of large amounts of purifiable proteins has already been a great asset, as has the development of techniques for crystallizing something other than globular proteins, such as membrane proteins (Michel, 1982). These developments, combined with greatly improved procedures for the collection of X-ray data (Xuong *et al*, 1978), make it altogether likely that a truly representative set of crystal structures can be gathered.

Even in those cases where your sequence resembles one from a protein whose crystal structure is not known, there is often much that can be learned. For example, March and Inouye (1985a) sequenced an open reading frame from *E. coli* that exists adjacent to the *lep* gene coding for the signal peptidase, but they were unable to assign a function to the gene product. A computer search revealed that it resembles the elongation factor family, members of which all bind GTP. In fact, when those investigators expressed their mystery protein in an amplified system, they were able to photolabel it with a GTP-derivative (March & Inouye, 1985b).

Table IX. Some Proteins That Contain Repeated Sequences

Protein (source)	Repeat	Repeat Length	Number Repeats
silk fibroin (silkworm)	GA	2	(multiple)
iron transport factor (E. coli)	PX	2	(multiple)
collagen (animals)	GPX or GXP	3	(multiple)
antifreeze proteins (fish)	AAT	3	(multiple)
malarial antigen (plasmodia)	NANP	4	(multiple)
keratin (vertebrates)	CCXPY	5	(multiple)
chorion proteins (inverts)	GYGGL	5	(multiple)
protamines (fish sperm)	ARRRR	5	6
myosins (animal)	-	7	(multiple)
apolipoproteins (mammals)	-	11	13-26
fibrinogen α chain (mammals)	-	13	6-10
salivary proteins (mammals)	-	14	17
fibrinogen α chain (lamprey)	-	18	>4
ribosomal protein S1 (E. coli)	-	22	12
EGF precursor (mouse)	(6-Cys)	40±2	10
complement factor H (mouse)	(4-Cys)	61	20

Not everyone is so lucky, however, and it may be that the search with your sequence didn't retrieve any significant matches. There are many other things you can do with the computer that can help you learn about your protein, however, so don't despair.

Internal Repeats

Many proteins contain internally repeating unit sequences. The length of the repeat can range from a lower limit of two (even one, if you want to think of some long stretches of the same residue that way) to internal duplications of the order of several hundred residues. Some idea of the range of such repeats can be derived from a consideration of Table IX. Repeated sequences frequently have structure-function connotations and should always be carefully weighed. Indeed, the elongation of proteins, in general, is achieved by tandem duplications. Some classic examples are clostridial ferredoxin, which at 55 residues includes two 26-residue repeats (Eck & Dayhoff, 1966), and *E. coli* β-galactosidase, the 1031-residue structure of which includes two 398-residue duplicons (Hood *et al*, 1980). A search for internal repeats that comes up empty may also be useful in estimating the time of invention. If a long sequence shows no sign of internal duplication, then it can be presumed that the protein is very ancient or is changing very rapidly. A simple species comparison should settle the point.

It should be understood that when the sequence comparer refers to a repeat, he seldom means an *exact* repeat. The implication is that the repeat was surely exact immediately upon its occurrence, but amino acid replacements begin eroding the correspondance straight-away. One of the nicest ways to show repeats in a sequence is with a "self-diagonal" plot (Figure 16). In this setting, a sequence is compared with itself, and if there are internal similarities they show up as off-set diagonals, in addition to the main diagonal that results from self-identity.

Consensus Sequences

There are many short sequences or patterns that are often (but not always) diagnostic of certain binding properties or active sites (Table X). These can be set into a small subcollection and searched against your sequence at a low level of stringency in a